



Psychometric evaluation of a statistical literacy instrument for pre-service mathematics teachers: Rasch analysis using R

Putri Nur Malasari*

Universitas Islam Negeri
Sunan Kudus, INDONESIA

Siti Qomariyah

Universitas Islam Negeri
Sunan Kudus, INDONESIA

Savira Najwa Meilani

Universitas Islam Negeri
Sunan Kudus, INDONESIA

Article Info

Article history:

Received: Feb 07, 2026

Revised: May 30, 2026

Accepted: August 22, 2026

Keywords:

Educational Measurement;
Pre-Service Mathematics
Teachers;
R Rasch Model;
Statistical Literacy.

Abstract

Background: Statistical literacy is an essential competency for pre-service mathematics teachers, characterized by persistent challenges in interpreting and critically evaluating statistical information. Despite the development of several statistical literacy instruments, comprehensive Rasch-based psychometric evidence for instruments targeting pre-service mathematics teachers remains limited.

Purpose: This study aimed to evaluate the psychometric quality of a statistical literacy instrument for pre-service mathematics teachers using the Rasch measurement model implemented in R. The evaluation focused on item functioning, reliability, validity, and overall measurement characteristics.

Methods: Data were collected from 132 pre-service mathematics teachers enrolled in an education program. Rasch analysis was conducted to examine item fit and difficulty, reliability, dimensionality, local independence, and person-item targeting.

Results: The entire 15 items met the Rasch fit criteria (Infit and Outfit MNSQ = 0.5–1.5), thereby supporting the construct validity and internal consistency, with no misfitting items identified. Most items ranged from easy to moderate (–1.47 to +0.35 logits), with only a particular difficult item. Wright Map analysis showed that respondents' abilities generally exceeded the highest item difficulty, suggesting suboptimal targeting and reduced measurement precision for high-ability respondents.

Conclusions: The results showed a mild ceiling effect, with 40.2% of respondents exceeding the hardest item and 3.0% achieving perfect scores. Generally, the instrument exhibited adequate psychometric quality but required more difficult items to improve measurement precision for high-ability respondents. The script-based Rasch workflow in R also enhanced the transparency and reproducibility of psychometric evaluation.

To cite this article: Malasari, P. N., Qomariyah, S. & Meilani, S. N. (2026). Psychometric evaluation of a statistical literacy instrument for pre-service mathematics teachers: Rasch analysis using R. *Journal of Advanced Sciences and Mathematics Education*, 6(3), 586-604.

INTRODUCTION

Statistical information is increasingly influencing public discourse, educational practice, and evidence-based decision making. As a result, concerns about the limited ability of individuals to critically evaluate data have intensified, underscoring the need for valid statistical literacy measures. The growing availability of digital data, and its use in daily context further strengthened the importance of statistical literacy as an essential competency for both students and teachers (Schreiter et al., 2024). Statistical literacy is widely conceptualized as the ability to critically interpret, evaluate, and communicate interrelated information for informed decision making (Berndt et al., 2021; Friedrich et al., 2024; Koga, 2022; Schüller, 2022; Umbach, 2022). Recent studies reported that it extends beyond procedural competence to include critical reasoning, evaluation of data-based

*Corresponding author:

Putri Nur Malasari, Universitas Islam Negeri Sunan Kudus, Indonesia
putrinurmalasari@uinsuku.ac.id ✉

claims, and responsible use of similar evidence in authentic contexts (Ozturk & Baki, 2024; Schreiter et al., 2024).

In line with the description above, statistical literacy has become a central objective of statistics education, enabling students to critically evaluate data-based arguments, interpret information presented in the media, as well as engage in evidence-informed decision making on societal issues. Several recent studies have outlined that statistical and data literacy are fundamental competencies for addressing contemporary challenges associated with data-rich environments and should be systematically developed throughout teachers education programs (Schreiter et al., 2024). The development process is crucial for both school students and pre-service mathematics teachers, as it is responsible for fostering diverse competencies in classrooms.

The main factor influencing students' opportunities to develop data-informed reasoning and critical engagement with related information is the statistical literacy of pre-service mathematics teachers. The possession of procedural knowledge alone is insufficient for effective statistics teaching. Furthermore, these teachers may struggle to foster students' statistical reasoning and informed decision making in increasingly data-rich learning environments, due to the inability to critically interpret data and evaluate statistical arguments (Schreiter et al., 2024). Recent evidence showed that statistical literacy is closely associated with the ability to design meaningful investigations, interpret authentic datasets, and facilitate students' reasoning, making it an essential component of mathematics teachers competence (Ozturk & Baki, 2024; Schreiter et al., 2024).

The persistent difficulties reported among pre-service teachers in interpreting statistical information, evaluating data-based claims, and determining higher-order statistical reasoning outlined the importance of obtaining valid evidence about statistical literacy competencies (Ozturk & Baki, 2024; Schreiter et al., 2024). This evidence depends on the availability of instruments with sound psychometric properties. Relatively few analyses have reported comprehensive Rasch-based evaluation of statistical literacy instruments developed for pre-service mathematics teachers, leaving questions regarding the quality and interpretability of the resulting measurement. The challenges documented outlined the need for instruments that provide accurate and meaningful measures of statistical literacy. The production of valid and reliable measurement evidence requires rigorous psychometric evaluation, limited to instruments designed for pre-service mathematics teachers. Recent evidence suggested that higher-order dimensions of statistical literacy remained difficult for many pre-service teachers to master. Carried out a study on 202 pre-service mathematics teachers, and reported that the advantages of statistics instruction were evident in basic and procedural aspects, while challenges persisted in data interpretation alongside more complex components. This is consistent with the result of, which identified continuing limitations in the statistical and data literacy competencies required for effective teaching. Majority of trainee teachers showed competence in foundational statistical literacy, with only a few exhibiting higher-order statistical reasoning, including the interpretation of complex data representations, critical evaluation of statistical evidence, and justification of conclusions based on data (Forgasz et al., 2024). The implications of the numerous challenges extend beyond teachers competence to the quality of statistics education. Previous studies reported that many teachers were inadequately prepared to teach statistics, particularly conceptually demanding topics such as variability, sampling distributions, p-values, and confidence intervals, despite acknowledging the growing relevance in contemporary society.

Significant methodological gaps remain in the validation of instruments developed for pre-service mathematics teachers, despite the growing interest in statistical literacy assessment. Existing evidence is still insufficient to prove whether these instruments produced valid, reliable, and interpretable measurement across multiple psychometric dimensions. In many published studies, evidence regarding item functioning, dimensionality, local independence, person-item targeting, and

measurement precision has been reported partially or omitted altogether. This has limited confidence in the resulting measurement interpretations (Bond et al., 2020).

Another unresolved issue is the transparency and reproducibility of psychometric validation. Recent developments in educational measurement have increasingly outlined the importance of open and reproducible analytical workflows to enhance the credibility, replicability, and independent verification of the results (Shimizu1 & Ferreira2, 2023). However, reproducible Rasch validation using open-source software remains uncommon in the development of statistical literacy instruments for pre-service mathematics teachers, resulting in difficulty for other studies to replicate analytical procedures or verify reported psychometric evidence.

Recent reviews have outlined the growing need for robust assessment instruments capable of accurately measuring statistical and data literacy competencies required in contemporary teachers education (Schreiter et al., 2024). Current evidence provides limited guidance regarding the capability of existing instruments to adequately represent these competencies, including simultaneously satisfying modern psychometric standards. Therefore, a methodological gap remains between the increasing demand for high-quality statistical literacy assessment and the availability of comprehensively validated, transparently analysed measurement instruments.

The present study addressed certain limitations by evaluating the measurement quality of a statistical literacy instrument for pre-service mathematics teachers. This was achieved using a comprehensive Rasch analysis implemented through a fully reproducible workflow in R. Furthermore, the comprehensive psychometric evidence supported statistical literacy assessment in mathematics teacher education by integrating evidence of validity, reliability, dimensionality, item functioning, and person-item targeting within a transparent analytical framework.

The present study aims to evaluate the measurement quality of a statistical literacy instrument for pre-service mathematics teachers using the Rasch model. The psychometric properties of the instrument, namely validity evidence, reliability, item difficulty, and person-item targeting, were examined by assessing the statistical literacy levels of prospective mathematics teachers. The provision of a transparent and reproducible Rasch-based validation process enabled the contribution of psychometric evidence to support the use and further development of statistical literacy assessment in mathematics teachers education.

LITERATURE REVIEW

The main limitation of Classical Test Theory (CTT) is its inability to estimate item characteristics and respondents' ability on a common measurement scale, thereby reducing the depth of information available for evaluating instrument quality (Rizbudiani et al., 2021). This method relies on total scores and does not provide detailed information regarding the quality of each test item (Bond et al., 2020). Rasch estimates are sample-independent when the model fits, supporting item invariance. It is dissimilar to CTT, because item statistics tend to depend on the sample. As an alternative, the Rasch model offered a more robust method to instrument analysis, considering the capability to model the relationship between individual ability and item difficulty independently. The model enabled less sample dependent measurement and provided important information, namely item fit, reliability, and unidimensionality of the instrument (Denovan et al., 2026; Pretorius, 2023; Ribeiro Santiago et al., 2020). The Rasch model also provides detailed information on item fit, reliability, and the disposition between respondents' ability and item difficulty levels. This provided evidence supporting the validity of the resulting instrument used for the ongoing evaluation of prospective mathematics teachers (Chue, 2022). The diverse indicators jointly established the ability of the items to define a single, replicable measurement scale. Therefore,

the use of the Rasch model is becoming increasingly relevant in the development of statistical literacy instruments for prospective mathematics teachers.

Advances in computing technology addressing reproducibility concerns that arise when analyses rely on closed, point-and-click software have driven the use of open-source software such as R in educational data reviews. A transparent, and thoroughly scripted analytical workflow offered important advantages for measurement studies by enhancing methodological transparency, facilitating reproducibility, and enabling the independent verification of analytical procedures. The open-source nature of R enabled the development of a wide range of analytical methods that are continually updated (Shimizu¹ & Ferreira², 2023). However, the use of R in Rasch analysis for the development of statistical literacy instruments remains scarce. This is because only a few published studies have applied a full Rasch workflow to statistical-literacy instruments for pre-service teachers using open-source tools, particularly within the context of teachers education.

The Rasch model was used to evaluate instruments, including the abilities of trainee teachers, particularly in the context of statistical literacy and other forms of literacy. For example, Guven et al. (2021) focused on using the model to analyse the level of statistical literacy among trainee mathematics teachers. It was reported that statistical reasoning, particularly variability, sampling, and inference were suboptimal. Other similar studies have also shown that the model was effective in testing the validity and reliability of educational instruments, including in the measurement of literacy and teachers competence (Boone et al., 2014). However, these studies did not combine a comprehensive Rasch evaluation with an open and reproducible workflow.

METHOD

Research Design

The psychometric properties of the statistical literacy instrument for pre-service mathematics teachers were examined using the Rasch measurement model. In addition, the analysis focused on item functioning, reliability, validity, and measurement precision to evaluate the overall quality of the instrument

Participant

A total of 132 sixth-semester pre-service mathematics teachers from a university in Kudus, Indonesia, participated in this study. The respondents were selected through purposive sampling based on the criterion that the teachers had completed or were currently enrolled in a data analysis and visualization course. The sample consisted of 87 female and 45 male pre-service mathematics teachers aged 19–21 years. That a sample of at least 100 respondents was sufficient to achieve stable item calibration in Rasch analysis. Therefore, the sample size was considered adequate for evaluating the psychometric properties of the instrument.

Population and the methods of sampling Instrumentation (sample of questions, scoring method, and psychometric properties (validity and reliability)):

Instrument

The instrument was developed by synthesizing and adapting significant dimensions of statistical literacy proposed by previous studies. The indicator of understanding data was mainly adapted from the frameworks designed by previous studies, while the presentation of data was mainly derived from previous studies. The communicating data indicator was informed by previous studies, with the interpretation process adapted from previous studies. These indicators were operationalized into 15 items designed to assess statistical literacy among the teachers. Prior to empirical testing, the items were reviewed by three experts in mathematics education and educational measurement. Content validity was evaluated using Aiken's V to examine the relevance, clarity, and representativeness of each item. Table 1 shows the statistical literacy indicators and operational definitions that guided the development and validation of the instrument.

Table 1. Statistical Literacy Indicators

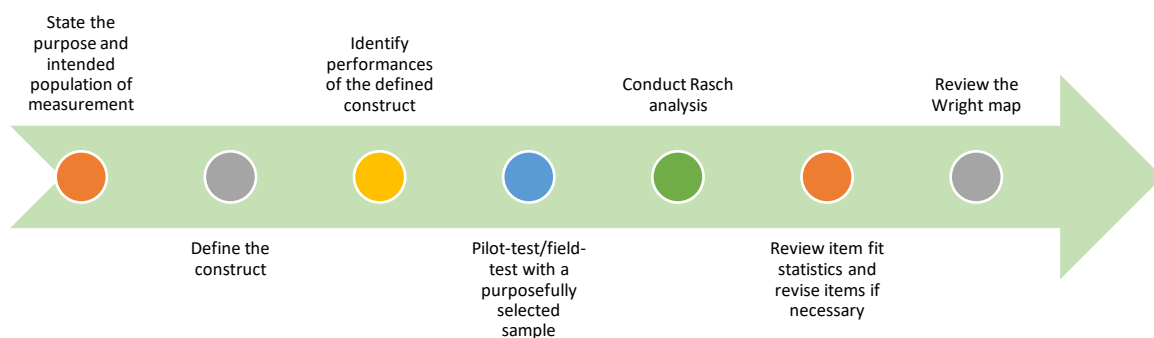
Indicators	Description
Understanding the Data	Prospective mathematics teachers are able to understand statistical concepts, such as interpreting various forms of data presented in tables, graphs, diagrams or other formats. They are also able to draw statistical conclusions from the data.
Presenting the Data	Prospective mathematics teachers can present data accurately using tables, diagrams, graphs or other means, and can record the key information contained in the data
Communicating the Data	Prospective mathematics teachers are able to communicate the process of statistical data analysis systematically through clear and structured writing
Interpreting the Data	Prospective mathematics teachers are able to interpret data accurately and identify statistical concepts that can be used to draw accurate statistical conclusions

The statistical literacy items were developed based on the indicators presented in Table 1 and were formatted as multiple-choice items with four response options. Responses were scored dichotomously (1 = correct, 0 = incorrect) and analysed using the Rasch measurement model to evaluate item functioning, person ability, reliability, and overall measurement quality.

Procedures and if relevant, the time frame:

Analysis plan (describe statistical tests and the comparisons made; ordinary statistical methods should be used without comment; advanced or unusual methods may require a literature citation)

The development of the statistical literacy instrument followed the framework. This framework was selected considering the integration of construct specification, item development, pilot testing, and Rasch-based psychometric evaluation within a systematic instrument design process. Liu's framework provided an appropriate methodological foundation for guiding instrument development and empirical validation, because the present study focused on evaluating the measurement quality of a statistical literacy instrument using the Rasch model. Based on this perspective, the stages of the framework adopted are shown in Figure 1.

**Figure 1.** Steps in the Development of Statistical Literacy Instruments

Prerequisite tests were carried out to ensure the data met the basic assumptions of Rasch modelling. This included the unidimensionality test, which mandated the instrument to measures only a single dimension of ability. Unidimensionality was assessed using a PCA of the standardized residuals. Subsequently, the Local Independence test was carried out to prove that responses to a particular item were not influenced by responses to the other. The Monotonicity test ensured that the probability of answering correctly was directly proportional to the ability of the respondent (the more intelligent the respondent, the greater the chances of answering correctly). Item validity was

measured through the fit of the data to the model (Infit and Outfit Mean Square), with the productive range within $0.5 \leq \text{MNSQ} \leq 1.5$.

In this context, an item fit analysis was conducted using the Infit and Outfit MNSQ statistics to assess how well each item conformed to the assumptions of the Rasch model. Infit and Outfit MNSQ were more sensitive to response patterns near a respondent's ability level and unexpected responses, respectively. These indices were collectively used to evaluate whether each item functioned consistently with the Rasch model.

The criteria for item acceptance are as follows:

Ideal range: $0.5 < \text{MNSQ} < 1.5$

A value of 1.0 indicates the expected (ideal) fit

A value > 1.5 indicates an underfit item (too unpredictable)

A value < 0.5 indicates an overfit item (too predictable)

Item difficulty estimates were generated to examine the distribution of item difficulty as well as assess the relationship between item difficulties and the ability levels of respondents. In addition, Table 2 shows the item difficulty classification used in this study.

Table 2. Item Difficulty Classification

Logit	Description
$< -2,0$ logit	Very easy
$-2,0$ until $-1,0$ logit	Easy
$-1,0$ until $0,0$ logit	Moderate
$> 0,0$ logit	Difficult

The reliability of the statistical literacy instrument was assessed, alongside the estimation of three indices. This included person (EAP/PV = 0.752; WLE = 0.671), and item separation reliabilities (0.855), as well as respective separation indices. In the context of the Rasch Model, reliability refers to the consistency and stability of the measurement. The variable can be evaluated from several aspects, such as person and item separation, including EAP/PV.

The distribution of prospective mathematics teachers' abilities was examined using a Wright map. In this context, the map was used to examine the disposition between respondent abilities and item difficulties. The results showed that respondent abilities were distributed across a wider range than item difficulties, suggesting suboptimal person item targeting. This visualisation allowed the evaluation of person-item targeting, which determined measurement precision and any ceiling effect.

RESULTS AND DISCUSSION

Results

Testing the Prerequisites and Basic Assumptions of the Rasch Model

1. Test for Unidimensionality

Figure 2 shows the R script used to assess the unidimensionality assumption. This was performed through residual principal component analysis (PCA) within the Rasch measurement framework.

```
# --- R: person estimates, difficulties, standardized residuals, residual PCA ---
wle <- tam.wle(mod)
theta <- wle$theta; perr <- wle$error
b <- mod$xi$xi; names(b) <- colnames(full_data); b_se <- mod$xi$se.xsi

Xc <- as.matrix(full_data)
P <- 1 / (1 + exp(-outer(theta, b, "-")))
resid <- Xc - P
std_resid <- resid / sqrt(P * (1 - P))
ev <- eigen(cor(std_resid), only.values = TRUE)$values # residual eigenvalues

cat("UNIDIMENSIONALITY\n")
cat(sprintf(" EAP/PV reliability : %.3f (>= 0.70 acceptable)\n", mod$EAP.rel))
cat(sprintf(" First residual contrast (eigenvalue): %.2f (< 2.0 supports unidimensionality)\n", ev[1]))
cat(sprintf(" Next residual eigenvalues : %.2f, %.2f, %.2f\n", ev[2], ev[3], ev[4]))
```

Figure 2. R Script for Assessing Unidimensionality Using PCA of Rasch Residuals

A residual principal component analysis PCA was conducted using the R script to assess the unidimensionality of the statistical literacy instrument. Based on this perspective, the results of the analysis are shown in Table 3.

Table 3. Results of the Rasch Unidimensionality Analysis

Parameter	Value	Criterion	Interpretation
EAP/PV Reliability	0.752	≥ 0.70	Acceptable reliability
First Residual Contrast Eigenvalue	1.75	< 2.00	Supports unidimensionality
Second Residual Eigenvalue	1.59	< 2.00	No substantial secondary dimension
Third Residual Eigenvalue	1.46	< 2.00	No substantial secondary dimension
Fourth Residual Eigenvalue	1.39	< 2.00	No substantial secondary dimension

As shown in Table 3, the EAP/PV reliability coefficient was 0.752, implying acceptable measurement reliability. In addition, the first residual contrast eigenvalue was 1.75, which was below the recommended threshold of 2.0. The remaining residual contrasts also produced eigenvalues less than 2.0. These results suggested the absence of a substantial secondary dimension in the residual structure and the predominant measurement of a single latent construct by the instrument.

2. Test for Local Independence

Following the assessment of unidimensionality, the assumption of local independence was evaluated using Yen's Q3 statistic. This procedure examined residual correlations among item pairs after controlling the latent trait, thereby identifying potential local dependencies between items. The methodological transparency and reproducibility were enhanced by ensuring the R script used to compute Yen's Q3 statistics, including the summarized local independence diagnostics are shown in Figure 3.

```
# --- R: Yen's Q3 + summary table + most-correlated pairs (the table reviewers asked for) ---
Q3 <- cor(resid); diag(Q3) <- NA
ut <- upper.tri(Q3)
maxQ3 <- max(abs(Q3), na.rm = TRUE)
meanQ3 <- mean(Q3[ut], na.rm = TRUE)
Q3star <- maxQ3 - meanQ3
n_flag <- sum(abs(Q3[ut]) > 0.20, na.rm = TRUE)

# (a) summary table
LI_summary <- data.frame(
  Statistic = c("Mean Q3 (off-diagonal)", "Max |Q3|", "Q3* (max - mean)",
    "Pairs with |Q3| > 0.20", "Total item pairs"),
  Value = c(sprintf("%.3f", meanQ3), sprintf("%.3f", maxQ3), sprintf("%.3f", Q3star),
    as.character(n_flag), as.character(choose(3, 2))),
  Criterion = c("-", "< 0.20", "< 0.20", "few / none", "-"),
  Status = c("-",
    ifelse(maxQ3 < 0.20, "Met", "Check"),
    ifelse(Q3star < 0.20, "Met", "Check"),
    ifelse(n_flag == 0, "Met", "Check"), "-")
)
cat("LOCAL INDEPENDENCE - Yen's Q3 summary\n"); print(LI_summary, row.names = FALSE)
write.csv(LI_summary, file.path(out_dir, "Table_LocalIndependence_Q3.csv"), row.names = FALSE)

# (b) the 8 item pairs with the largest |Q3|
idx <- which(ut, arr.ind = TRUE)
pairs_df <- data.frame(
  Item_A = colnames(full_data)[idx[, 1]],
  Item_B = colnames(full_data)[idx[, 2]],
  Q3 = round(Q3[ut], 3)
)
pairs_df <- pairs_df[order(-abs(pairs_df$Q3)), 1:min(8, nrow(pairs_df)), ]
cat("\nTop item pairs by |Q3|:\n"); print(pairs_df, row.names = FALSE)
write.csv(pairs_df, file.path(out_dir, "Table_LocalIndependence_TopPairs.csv"), row.names = FALSE)
```

Figure 3. R Script for Assessing Local Independence Using Yen's Q3 Statistic

The residual correlations among item pairs were examined to identify potential local dependencies. Meanwhile, the results of the Yen's Q3 analysis are shown in Table 4.

Table 4. Results of the Local Independence Analysis Using Yen's Q3 Statistic

Panel A. Summary Statistics

Statistic	Value	Criterion	Interpretation
Mean Q3 (off-diagonal)	-0.069	-	-

Statistic	Value	Criterion	Interpretation
Maximum	Q3		0.286
Q3* (Max – Mean)	0.355	< 0.20	Some local dependence detected
Item pairs with	Q3	> 0.20	12
Total item pairs	105	–	–

Panel B. Item Pairs with the Highest Residual Correlations

Item A	Item B	Q3
Item 2	Item 14	-0.286
Item 1	Item 3	-0.277
Item 5	Item 6	-0.266
Item 6	Item 15	-0.262
Item 4	Item 11	0.253
Item 2	Item 7	-0.246
Item 11	Item 13	-0.241
Item 14	Item 15	-0.236

The mean residual correlation (Q3) was -0.069 , with the maximum absolute Q3 value (0.286) and adjusted Q3 statistic ($Q3^* = 0.355$) exceeding the recommended threshold of 0.20. This suggested the presence of some residual dependence among several item pairs. A total of 12 out of 105 item pairs showed residual correlations greater than $|0.20|$. The strongest residual associations were observed between Items 2 and 14 ($Q3 = -0.286$), 1 and 3 ($Q3 = -0.277$), alongside 5 and 6 ($Q3 = -0.266$). Although these results suggested some degree of local dependency, the affected item pairs represented a relatively small proportion of all possible item pairs. Therefore, the local independence assumption appears to be largely satisfied, with several item pairs warranting further examination.

3. Test for Monotonicity

The monotonicity assumption was examined using rest-corrected point-biserial correlations in line with the evaluation of local independence. This analysis assessed whether the probability of answering an item correctly increased as respondents' levels of statistical literacy rose. The R script used to conduct the monotonicity analysis is shown in Figure 4.

```
# --- R: rest-corrected point-biserial ---
rest <- rowSums(Xc)
ptbis <- sapply(1:J, function(j) cor(Xc[, j], rest - Xc[, j]))
names(ptbis) <- colnames(full_data)
print(data.frame(Item = names(ptbis), Point_Biserial = round(ptbis, 3),
  Direction = ifelse(ptbis > 0, "positive", "NEGATIVE")), row.names = FALSE)
cat(sprintf("\n Mean: %.3f | range: %.3f to %.3f | all positive: %s\n",
  mean(ptbis), min(ptbis), max(ptbis), all(ptbis > 0)))
```

Figure 4. R Script for Assessing Monotonicity Using Rest-Corrected Point-Biserial Correlations

The rest-corrected point-biserial correlations were calculated for all items to assess monotonicity and item discrimination. Moreover, the results of the analysis are shown in Table 5.

Table 5. Results of the Monotonicity Analysis Based on Rest-Corrected Point-Biserial Correlations

Item	Point-Biserial Correlation	Direction
Item 1	0.412	Positive
Item 2	0.538	Positive
Item 3	0.441	Positive
Item 4	0.432	Positive
Item 5	0.307	Positive
Item 6	0.274	Positive
Item 7	0.321	Positive
Item 8	0.382	Positive

Item	Point-Biserial Correlation	Direction
Item 9	0.519	Positive
Item 10	0.288	Positive
Item 11	0.558	Positive
Item 12	0.416	Positive
Item 13	0.359	Positive
Item 14	0.079	Positive
Item 15	0.515	Positive
Mean: 0.389 Range: 0.079–0.558 all positive: TRUE		

All items showed positive rest-corrected point-biserial correlations, suggesting that respondents with higher levels of statistical literacy tended to answer the items correctly. The point-biserial coefficients ranged from 0.079 to 0.558, with a mean value of 0.389. These results provided evidence that the items functioned in the expected direction and satisfied the monotonicity assumption. Furthermore, the positive correlations suggested that all items showed acceptable discrimination with respect to the underlying statistical literacy construct.

Item fit and Empirical Validity (Infit/Outfit)

Item fit analysis was conducted using the Rasch model to evaluate the empirical validity of the instrument. Infit and Outfit Mean Square (MNSQ) statistics were also examined to identify items that deviated from model expectations. This included determining whether each item contributed appropriately to the measurement of statistical literacy. In addition, the R script used for the item fit analysis is shown in Figure 5.

```
# --- R: Infit/Outfit + master item table ---
fit <- tam.fit(mod)$itemfit
item_table <- data.frame(
  Item = colnames(full_data), Difficulty = round(b, 2), SE = round(b_se, 2),
  Infit_MNSQ = round(fit$Infit, 2), Outfit_MNSQ = round(fit$Outfit, 2), PtBis = round(ptbis, 2))
within <- function(x) x >= 0.5 & x <= 1.5
item_tablesFit_Status <- ifelse(within(item_tables$Infit_MNSQ) & within(item_tables$Outfit_MNSQ), "Fit", "Misfit")
item_tablesNote <- ifelse(item_tables$Infit_MNSQ < 0.5 | item_tables$Outfit_MNSQ < 0.5, "overfit (low discrimination)"
  ifelse(item_tables$Infit_MNSQ > 1.5 | item_tables$Outfit_MNSQ > 1.5, "underfit (noisy)", "-"))
print(item_table, row.names = FALSE)
write.csv(item_table, file.path(out_dir, "Table1_ItemFit_Difficulty.csv"), row.names = FALSE)
cat(sprintf("\n Items within 0.5-1.5 on BOTH Infit & Outfit: %d of %d\n", sum(item_tables$Fit_Status == "Fit"), J))
```

Figure 5. R Script for Assessing Item Fit Using Infit and Outfit MNSQ Statistics

The item fit statistics were calculated to evaluate the empirical validity of the statistical literacy instrument. In line with the evaluation process conducted, the results of the analysis are shown in Table 6.

Table 6. Item Fit Statistics and Empirical Validity Evidence

Item	Difficulty (logit)	SE	Infit MNSQ	Outfit MNSQ	Point-Biserial	Fit Status
1	-1.18	0.21	0.97	0.95	0.41	Fit
2	-0.42	0.20	0.87	0.82	0.54	Fit
3	-1.22	0.22	0.96	0.84	0.44	Fit
4	-0.38	0.20	0.96	0.98	0.43	Fit
5	-0.87	0.21	1.07	1.17	0.31	Fit
6	-0.50	0.20	1.11	1.15	0.27	Fit
7	-1.32	0.22	1.04	1.13	0.32	Fit
8	-1.32	0.22	0.99	0.95	0.38	Fit
9	-1.32	0.22	0.88	0.81	0.52	Fit
10	-1.09	0.21	1.08	1.18	0.29	Fit
11	-1.47	0.23	0.84	0.66	0.56	Fit
12	-1.32	0.22	0.96	0.85	0.42	Fit
13	0.35	0.20	1.00	1.02	0.36	Fit
14	0.00	0.20	1.31	1.41	0.08	Fit
15	-0.74	0.20	0.89	0.82	0.51	Fit

Table 6 shows that the 15 items has Infit and Outfit MNSQ values within the recommended range of 0.50 to 1.50, reflecting acceptable fit to the Rasch model. The Infit and Outfit MNSQ values

ranged from 0.84 to 1.31 and 0.66 to 1.41, respectively. These results suggested that none of the items showed substantial underfit or overfit, reflecting consistency between the observed responses and model expectations. In addition, all items produced positive point-biserial correlations, ranging from 0.08 to 0.56. The positive point-biserial coefficients implied respondents with higher levels of statistical literacy answered the items correctly, providing evidence that the items functioned in the expected direction. Considering that Item 14 produced the lowest point-biserial correlation (0.08), its Infit and Outfit statistics remained within the acceptable range. Generally, all 15 items satisfied the Rasch model fit criteria and showed acceptable empirical validity.

1. Item Fit Diagnostic Plot

To complement the item fit statistics presented in Table 6, a graphical diagnostic analysis was conducted to visualize the relationship between Infit and Outfit Mean Square (MNSQ) values for each item. The R script used to generate the item-fit diagnostic plot is presented in Figure 6.

```
#
# BLOCK A - Item Fit Diagnostic Plot (paper style: green "SAFE ZONE" box,
# red dashed 0.5/1.5 borders, dark-green points, labels below)
#
options(repr.plot.width = 9, repr.plot.height = 8.5)
library(ggplot2)

p_fit <- ggplot(item_table, aes(x = Infit_MNSQ, y = Outfit_MNSQ)) +
  # safe-fit zone
  annotate("rect", xmin = 0.5, xmax = 1.5, ymin = 0.5, ymax = 1.5,
    fill = "#C8E6C9", alpha = 0.50) +
  annotate("text", x = 1.0, y = 1.0, label = "SAFE ZONE\n(Fit)",
    colour = "#A5D6A7", fontface = "bold", size = 7) +
  # critical boundaries (red dashed)
  geom_hline(yintercept = c(0.5, 1.5), linetype = "dashed",
    colour = "#EF5350", linewidth = 0.6) +
  geom_vline(xintercept = c(0.5, 1.5), linetype = "dashed",
    colour = "#EF5350", linewidth = 0.6) +
  # items
  geom_point(aes(colour = Fit_Status), size = 3) +
  geom_text(aes(label = Item), vjust = 1.9, size = 2.8, colour = "grey15") +
  scale_colour_manual(values = c("Fit" = "#2E7D32", "Misfit" = "#C62828"),
    name = "Status") +
  coord_cartesian(xlim = c(0.5, 1.5), ylim = c(0.5, 1.5)) +
  labs(title = "Item Fit Diagnostic Plot",
    subtitle = "Distribution of Infit vs Outfit MNSQ (Critical Range: 0.5 - 1.5)",
    x = "Infit MNSQ", y = "Outfit MNSQ") +
  theme_minimal(base_size = 13) +
  theme(plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(colour = "grey30"))

ggsave(file.path(out_dir, "Fig1_ItemFit_Diagnostic.png"),
  p_fit, width = 9, height = 8.5, dpi = 200)
print(p_fit)
```

Figure 6. R Script Used to Visualize Item Fit Based on Infit and Outfit MNSQ Statistics

The script was executed to produce the item-fit diagnostic plot. Additionally, the resulting visualization is shown in Figure 7.

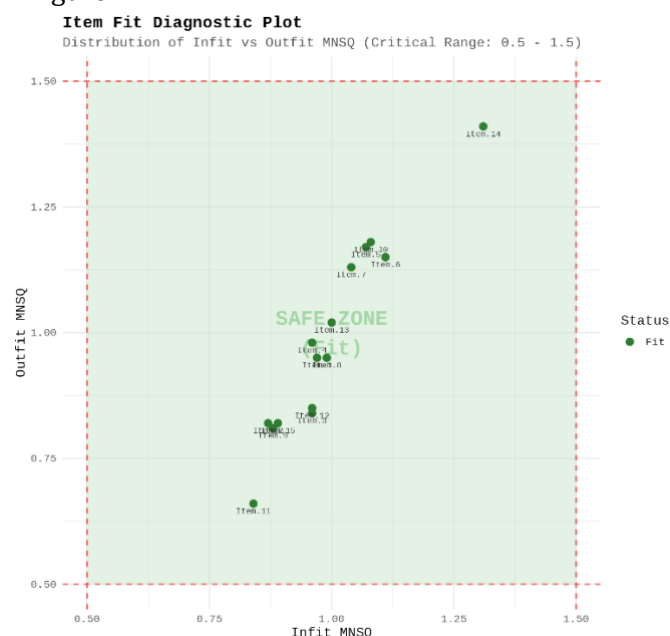


Figure 7. Item-Fit Diagnostic Plot Based on Infit and Outfit MNSQ Statistics

All items were located within the acceptable fit region defined by the Infit and Outfit MNSQ criteria (0.50–1.50). Meanwhile, none of the items showed substantial underfit or overfit, suggesting good agreement between the observed responses and Rasch model expectations. These results are consistent with the results shown in Table 6 and provide additional evidence supporting the empirical validity of the statistical literacy instrument.

Item Difficulty Analysis

Rasch difficulty estimates were calculated and categorized according to respective logit values to investigate the distribution of item difficulty. Figure 8 shows the R script used to perform the item difficulty analysis.

```
# --- R: difficulty classification ---
b2 <- item_table$Difficulty
cls <- cut(b2, breaks = c(-Inf, -2, -1, 0, Inf),
          labels = c("Very easy (< -2.0)", "Easy (-2.0 to -1.0)", "Moderate (-1.0 to 0.0)", "Difficult (> 0.0)"))
diff_tab <- data.frame(Item = item_table$Item, Difficulty = b2, Category = cls)
diff_tab <- diff_tab[order(diff_tab$Difficulty), ]
print(diff_tab, row.names = FALSE)
cat("\nCount per category:\n"); print(table(cls))
write.csv(diff_tab, file.path(out_dir, "Table2_ItemDifficulty.csv"), row.names = FALSE)
cat(sprintf("\n Difficulty range: %.2f to %.2f logits (spread %.2f)\n", min(b2), max(b2), max(b2) - min(b2)))
```

Figure 8. R Script for Item Difficulty Analysis and Classification

The script was executed to estimate item difficulty and perform classification according to respective logit values. Table 7 shows the resulting difficulty estimates and classifications.

Table 7. Item Difficulty Estimates and Difficulty Categories

Item	Difficulty (Logit)	Category
Item 11	-1.47	Easy
Item 7	-1.32	Easy
Item 8	-1.32	Easy
Item 9	-1.32	Easy
Item 12	-1.32	Easy
Item 3	-1.22	Easy
Item 1	-1.18	Easy
Item 10	-1.09	Easy
Item 5	-0.87	Moderate
Item 15	-0.74	Moderate
Item 6	-0.50	Moderate
Item 2	-0.42	Moderate
Item 4	-0.38	Moderate
Item 14	0.00	Moderate
Item 13	0.35	Difficult

The item difficulty estimates ranged from –1.47 to 0.35 logits, producing a difficulty spread of 1.82 logits. Most items were classified as easy (53.3%) or moderate (40.0%), while a particular item (Item 13) was categorized as difficult. None of the items fell within the very easy category. These results suggested that the instrument was mainly composed of items targeting low to moderate levels of statistical literacy, with limited representation of highly challenging items. The observed range of item difficulties showed that the instrument captured varying levels of respondent ability.

Reliability and Separation Analysis

Following the evaluation of item difficulty, reliability and separation analyses were conducted to assess the consistency and discriminatory power of the statistical literacy instrument. In the Rasch model, reliability coefficients suggest the reproducibility of person and item estimates.

Meanwhile, separation indices reflected the capability of the instrument to distinguish between different levels of respondent ability and item difficulty. The R script used to calculate reliability and separation statistics is shown in Figure 9.

```
# --- R: person & item separation ---
person_rel_wle <- 1 - mean(perr^2) / var(theta)
G_person <- sqrt(person_rel_wle / (1 - person_rel_wle)); strata_person <- (4 * G_person + 1) / 3
item_obs_var <- var(b); item_mse <- mean(b_se^2)
item_rel <- (item_obs_var - item_mse) / item_obs_var
G_item <- sqrt(item_rel / (1 - item_rel)); strata_item <- (4 * G_item + 1) / 3

rel_tab <- data.frame(
  Index = c("Person reliability (EAP/PV)", "Person separation reliability (WLE)",
    "Person separation index (G)", "Person strata",
    "Item separation reliability", "Item separation index (G)", "Item strata"),
  Value = round(c(mod$EAP.rel, person_rel_wle, G_person, strata_person, item_rel, G_item, strata_item), 3))
print(rel_tab, row.names = FALSE)
write.csv(rel_tab, file.path(out_dir, "Table3_Reliability.csv"), row.names = FALSE)
```

Figure 9. R Script for Reliability and Separation Analysis

The script was executed to estimate person reliability, item reliability, separation indices, and strata values. In addition, the results of the analysis carried out are shown in Table 8.

Table 8. Reliability and Separation Statistics

Index	Value
Person Reliability (EAP/PV)	0.752
Person Separation Reliability (WLE)	0.671
Person Separation Index (G)	1.428
Person Strata	2.238
Item Separation Reliability	0.855
Item Separation Index (G)	2.433
Item Strata	3.577

The person reliability coefficient based on EAP/PV estimates was 0.752, showing acceptable reliability in measuring the statistical literacy of respondents. The person separation reliability following WLE estimates was 0.671, suggesting a moderate level of consistency in distinguishing respondents according to respective ability levels. The person separation index ($G = 1.428$) and strata value (2.238) imply that the instrument was capable of differentiating respondents into two distinct levels of statistical literacy. Although the separation was adequate, the relatively modest separation index suggested that the instrument may have limited capability to discriminate among respondents with similar ability levels. The item separation reliability was 0.855, reflecting good precision in estimating item difficulty. The item separation index ($G = 2.433$) and strata value (3.577) suggested that the instrument successfully distinguished items across approximately three to four levels of difficulty. These showed that the item difficulty hierarchy was properly established, and the instrument reflected satisfactory measurement quality within the Rasch model.

Student Proficiency Distribution and Person Item Targeting

The reliability and separation analysis enabled the examination of student proficiency distribution and person-item targeting to evaluate the disposition between respondents' ability levels and item difficulty. In the Rasch model, effective targeting was achieved when the distribution of item difficulties corresponded closely to the distribution of respondent abilities. Moreover, the R script used to estimate person proficiency and assess person-item targeting is shown in Figure 9.

```
# -----
# BLOCK B — Wright Map (paper style: blue violin + jittered students on the
# left, green diamonds + labelled items on the right, dashed 0 line)
# -----
options(repr.plot.width = 10, repr.plot.height = 8)
library(ggplot2)
library(ggrepel)
person_df <- data.frame(grp = "STUDENTS\n(High Ability at Top)", logit = theta)
item_df <- data.frame(grp = "ITEMS\n(Hardest at Top)",
  logit = b, Item = names(b), Status = "Fit")
wm <- ggplot() +
  geom_violin(data = person_df, aes(x = grp, y = logit),
    fill = "#D6E4F0", colour = NA, alpha = 0.80, width = 1.0) +
  geom_jitter(data = person_df, aes(x = grp, y = logit),
    width = 0.13, colour = "#588DB8", alpha = 0.60, size = 1.7) +
  geom_point(data = item_df, aes(x = grp, y = logit, colour = Status),
    shape = 18, size = 4) +
  geom_text_repel(data = item_df, aes(x = grp, y = logit, label = Item),
    nudge_x = 0.30, direction = "y", hjust = 0, size = 3.1,
    segment.colour = "grey55", segment.size = 0.3,
    box.padding = 0.15, max.overlaps = 20) +
  geom_hline(yintercept = 0, linetype = "dashed", colour = "grey40") +
  scale_colour_manual(values = c("Fit" = "#2E8B57"), name = "Status") +
  labs(title = "Wright Map: Person Ability vs Item Difficulty",
    subtitle = "Logit Hierarchy Visualization (Combined Data N=132)",
    x = NULL, y = "Logit Scale (Measure)") +
  theme_minimal(base_size = 13) +
  theme(plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(colour = "grey30"),
    legend.position = "bottom",
    panel.grid.major.x = element_blank())
ggsave(file.path(out_dir, "Fig2_WrightMap.png"),
  wm, width = 10, height = 8, dpi = 200)
print(wm)
```

Figure 9. R Script for Student Proficiency Distribution and Person–Item Targeting Analysis

The script was executed to estimate respondent ability distributions and examine the disposition between person abilities and item difficulties. Meanwhile, the resulting person–item targeting results are shown in Figure 10.

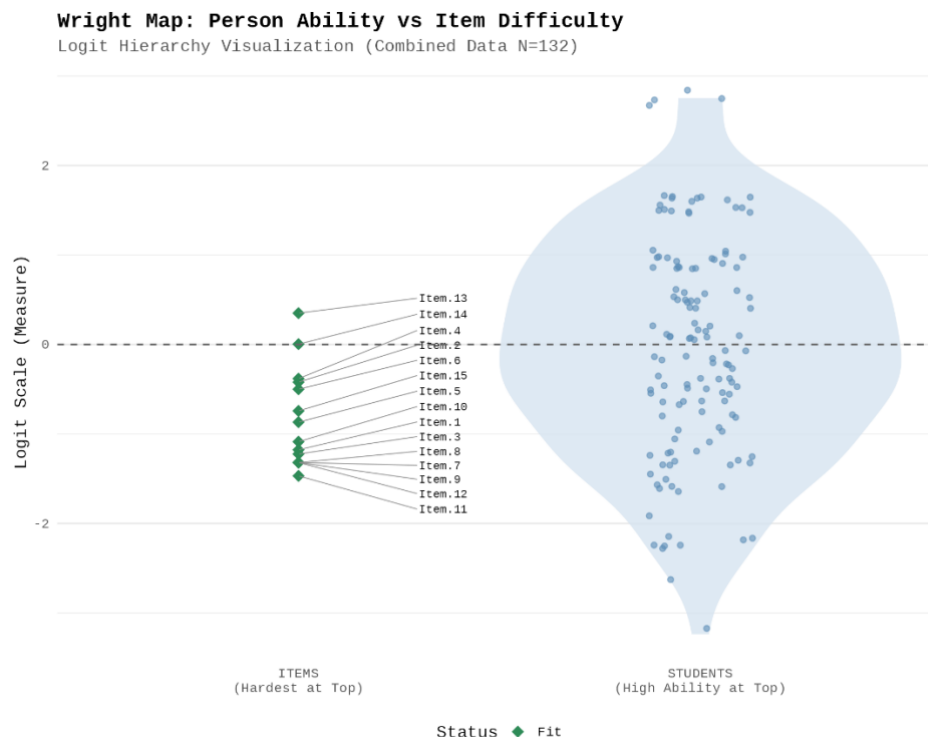


Figure 10. Wright Map of Person Ability and Item Difficulty Distribution

The Wright Map shows the distribution of respondent abilities and item difficulties on a common logit scale. The results show that respondent abilities were distributed over a wider range than item difficulties. Meanwhile, the person measures and item difficulties ranged from approximately –3 to +3 logits and –1.47 to 0.35 logits, respectively. Most items were located below

the average ability level of the respondents, reflecting that the instrument was easy for the sample under investigation. Items 13 and 11 were identified as the most difficult and easiest items, respectively. However, even the most difficult item remained substantially below the ability levels of several respondents, suggesting that the instrument contained relatively few challenging items. These results showed that the instrument was reasonably targeted toward respondents with low to moderate levels of statistical literacy and was less effective in measuring individuals with higher ability levels. Future studies should incorporate more difficult items to improve targeting across the full range of respondent abilities.

Discussion

This study provided evidence that the instrument developed possessed satisfactory psychometric properties for measuring statistical literacy among trainee mathematics teachers. The results suggested acceptable psychometric quality was not limited solely to items that fit the model. Although each item met the Rasch fit criteria, the instrument was found to measure respondents with low to moderate levels of statistical literacy more accurately compared to higher levels of ability. This distinction is significant because Rasch measurement outlined the dependence of high-quality assessment on item fit, including the disposition between item difficulty and respondent ability. Therefore, an instrument may meet model expectations and still provide limited information for specific segments within the ability range (Bond et al., 2020).

The results of the satisfactory item scale fit showed that each item consistently contributed to the measurement of the latent construct in question. This is consistent with recent studies on the development of Rasch-based instruments, including teachers education and literacy assessment. For example, reported satisfactory fit statistics provided evidence that the items collectively represented a single construct and produced stable measurement (Allen et al., 2022; Hudson et al., 2020; Shaw et al., 2021; Sinclair et al., 2021). Further stated acceptable fit statistics should be interpreted as evidence that the measurement model accurately explained respondents' interactions with the items, rather than confirming statistical adequacy (Nima et al., 2020; Tesio et al., 2024). Generally, these studies reinforced the view that Rasch fit statistics provided evidence supporting construct validity by showing consistency between theoretical expectations and observed response patterns. The recent psychometric literature also outlined that satisfactory fit should not be regarded as sufficient evidence of overall measurement quality, as validity and reliability remain equally important components of effective assessment.

More significant findings relate to the mismatch between respondents' ability and the difficulty level of the items. The Wright map shows that the majority of respondents demonstrated ability exceeding the difficulty level of the most difficult items, indicating that the instrument provides relatively limited information for distinguishing between highly proficient respondents. Similar limitations in person-item fit have been reported in recent Rasch-based validation studies. For example, in validating the Scientific Inquiry Literacy Instrument, emphasised that Wright map analysis is crucial as it reveals whether an instrument adequately targets the intended respondents, going beyond what item fit statistics alone can demonstrate (Darman et al., 2024; Liu et al., 2024; Shahat et al., 2024). Similarly, the application of recent Rasch analysis in educational measurement has shown that acceptable item fit and reliability can coexist with sub-optimal targeting when item difficulty levels are concentrated within a narrow range of ability, thereby reducing measurement precision for respondents at the upper end of the latent trait continuum (Engelhard et al., 2024). Consequently, this instrument functions effectively as a diagnostic tool for identifying respondents with low or moderate statistical literacy, but becomes increasingly less accurate when measuring advanced-level competence.

One possible explanation for these findings relates to the multidimensional nature of statistical literacy itself. Contemporary conceptualisations describe statistical literacy as extending beyond simply reading graphs or performing statistical calculations; it also encompasses the critical interpretation of data, the evaluation of statistical arguments, the recognition of misleading information, and evidence-based decision-making. These higher-order competencies generally place far greater cognitive demands than basic data interpretation tasks. As this instrument was essentially developed to ensure a broad representation of the fundamental indicators relatively few items require respondents to engage in complex statistical reasoning. Consequently, respondents with stronger statistical literacy were able to answer the majority of items correctly, resulting in the observed ceiling effect. This interpretation is consistent with recent studies on statistical literacy among trainee teachers, which recommend incorporating authentic contexts, conflicting statistical evidence, and real-world decision-making scenarios to better distinguish respondents at higher proficiency levels.

The findings of this study also contribute to the development of research on the assessment of statistical literacy in teacher education. Whilst previous research has largely examined the levels of statistical literacy among trainee teachers or evaluated learning interventions to improve these competencies (Schreiter et al., 2024), relatively few studies have focused on the psychometric quality of the instruments used to generate this evidence. This study expands this line of research by demonstrating that evidence of adequate item fit and reliability must be interpreted in conjunction with item-specific analysis when evaluating the quality of statistical literacy instruments. These findings are consistent with recent Rasch-based validation studies, which emphasise that robust instrument validation requires a variety of complementary sources of psychometric evidence, rather than relying solely on a single indicator of measurement quality (Darman et al., 2024; Engelhard et al., 2024). Therefore, this study not only contributes an empirically validated instrument but also provides evidence regarding its strengths and limitations in assessing statistical literacy amongst trainee mathematics teachers

From a practical perspective, these findings have important implications for mathematics teacher education. Statistical literacy is increasingly recognised as an essential professional competence, as trainee mathematics teachers are expected to interpret data critically, evaluate statistical claims presented in the media, and guide pupils in evidence-based reasoning. Recent international studies also argue that assessment instruments should measure higher-order statistical reasoning, rather than focusing solely on procedural knowledge or basic data interpretation (Forgasz et al., 2024; Schreiter et al., 2024). Consequently, the current instrument is suitable for diagnostic and formative assessment of pupils with low to moderate levels of statistical literacy. However, the observed limitations in targeting suggest that additional high-difficulty items are required before this instrument can be used with confidence to distinguish respondents with advanced statistical literacy. Incorporating authentic data analysis tasks, critical evaluation of statistical arguments, and the interpretation of complex graphical information is likely to improve measurement accuracy at the upper end of the ability continuum, in line with recent recommendations for literacy assessment in teacher education (Darman et al., 2024).

Overall, these findings suggest that the evaluation of statistical literacy instruments requires consideration of various complementary aspects of measurement quality. Although satisfactory item validity and reliability indicate that the instrument functions consistently within the Rasch measurement framework, the Wright plot reveals that measurement precision is reduced for respondents with relatively high statistical literacy because the available items do not adequately target that range of ability. These results highlight the importance of integrating evidence from item fit, reliability, dimensionality, and person-item targeting to obtain a comprehensive evaluation of instrument quality, as recommended in contemporary Rasch measurement research (Bond et al.,

2020; Engelhard et al., 2024). Therefore, future research should expand the item pool by including more cognitively challenging statistical literacy tasks and conduct cross-validation of the revised instrument with a more diverse sample of prospective mathematics teachers from various educational contexts. Such efforts will strengthen both the generalisability of the instrument and its utility for educational research and practice.

Implications

The numerous results suggested that the developed instrument provided a valid and reliable measure for assessing the statistical literacy of pre-service mathematics teachers with low-to-intermediate ability levels. Therefore, the instrument could be used as an initial screening tool to identify students who require additional support in developing statistical literacy competencies. The instrument is also more appropriate for formative assessment and monitoring learning progress than for differentiating among high performing respondents, as the person item targeting analysis limited measurement precision at the upper ability levels. Future refinement of the instrument should focus on developing additional high difficulty items to improve its ability to discriminate among respondents with higher levels of statistical literacy. This study broadly showed the value of comprehensive Rasch validation in providing robust psychometric evidence to support the development and use of statistical literacy assessment instruments in mathematics teachers education.

Limitations and Suggestions for Future Research

Several limitations were acknowledged, despite the satisfactory psychometric properties shown in this study. The instrument was sensitive to measuring low to moderate levels of statistical literacy and had limited precision for respondents with higher ability levels. The person item map showed a mismatch between the distribution of item difficulty and respondent ability, particularly at the upper end of the ability continuum. Although the instrument could be used effectively to assess prospective mathematics teachers with low to intermediate levels of statistical literacy, it was less suitable for accurately identifying and differentiating highly proficient respondents. The results further provided clear directions for carrying out future studies. To improve person item targeting and measurement precision across the full range of statistical literacy abilities, the item pool should be expanded by developing and validating additional high-difficulty items, particularly within the range of approximately +0.5 to +2.0 logits. These items were expected to reduce the observed ceiling effect and enhance the ability of the instrument to discriminate among respondents with higher levels of statistical literacy.

In this context, Item 11 was identified as the easiest item (-1.47 logits) and may be reviewed to determine the ability to contribute sufficient measurement information in future versions of the instrument. The revision should be in line with further empirical evidence rather than immediate removal, because the item remained within the acceptable fit range (Outfit MNSQ = 0.66) and showed adequate discrimination (point-biserial = 0.56). Finally, more advanced psychometric analyses, such as Differential Item Functioning (DIF) should be carried out to examine whether individual items functioned equivalently across different respondent groups. The numerous analyses would provide additional evidence regarding the fairness and generalizability of the instrument, including further strengthening its application in mathematics teachers education.

CONCLUSION

The statistical literacy instrument showed evidence of validity according to Rasch indicators. This included item fit (Infit and Outfit MNSQ within 0.5–1.5), person and item separation reliability, and unidimensionality (initial residual contrast = 1.75). The 15 items of the statistical literacy scale showed an adequate fit with the Rasch model. In addition, the MNSQ Infit and Outfit values fell within

the acceptable range (0.5–1.5), providing evidence supporting construct consistency and measurement reliability. The instrument was dominated by easy-to-moderate items, poorly targeted for high ability respondents. The item difficulty ranged from -1.47 to $+0.35$ logits (a spread of 1.82 logits), with 53 per cent of items (8 of 15) falling below -1.0 logit, and a particular item (Item 13, $+0.35$ logits) marginally above 0 logit. Student ability ranged from approximately -2.5 to $+3.0$ logits, with a mean of -0.03 logits, including the concentration of most respondents at the centre of the scale. The item person map showed that a large proportion of prospective mathematics teachers possessed ability levels above the most difficult item. This result suggested a substantial discrepancy between the distribution of item difficulty and respondent ability. The gap of approximately 0.5–2.0 logits between the most difficult items and the abilities of many respondents suggested the presence of a ceiling effect, which resulted in limited measurement precision at the upper end of the ability continuum. Therefore, the instrument is more effective for assessing respondents with low to moderate levels of statistical literacy compared to higher ability levels. Future refinement of the instrument should focus on developing additional high difficulty items to improve targeting and measurement precision for high-ability respondents.

AUTHOR CONTRIBUTIONS STATEMENT

Putri Nur Malasari contributed to conceptualization, methodology, instrument development, formal analysis, data curation, visualization, writing the original draft, reviewing and editing the manuscript, and supervision. Siti Qomariyah contributed to methodology, validation, investigation, data collection, formal analysis, writing the original draft, and reviewing and editing the manuscript. Savira Najwa Meilani contributed to investigation, data collection, data curation, visualization, literature review, writing the original draft, and reviewing and editing the manuscript. All authors contributed to the interpretation of the findings, critically reviewed the manuscript, and approved the final version for publication.

REFERENCES

- Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science. *European Journal of Psychological Assessment*, 38(1), 1–5. <https://doi.org/10.1027/1015-5759/a000699>
- Berndt, M., Schmidt, F. M., Sailer, M., Fischer, F., Fischer, M. R., & Zottmann, J. M. (2021). Investigating statistical literacy and scientific reasoning & argumentation in medical-, social sciences-, and economics students. *Learning and Individual Differences*, 86, 101963. <https://doi.org/10.1016/j.lindif.2020.101963>
- Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. <https://doi.org/10.4324/9780429030499>
- Chue, K. L. (2022). Using the Rasch model to examine pre-service teachers approach to learning profiles. *Methodological Innovations*, 15(1), 86–95. <https://doi.org/10.1177/20597991221077908>
- Darman, D. R., Suhandi, A., Kaniawati, I., Samsudin, A., & Wibowo, F. C. (2024). Development and validation of scientific inquiry literacy instrument (SILI) using Rasch measurement model. *Education Sciences*, 14(3), 322. <https://doi.org/10.3390/educsci14030322>
- Denovan, A., Dagnall, N., Peverell, M., Drinkwater, K. G., Escolà-Gascón, Á., & Carter, D. (2026). Evaluating the psychometric properties of the Generic Conspiracist Beliefs Scale: A Rasch modelling approach. *Frontiers in Psychology*, 17, 1721668. <https://doi.org/10.3389/fpsyg.2026.1721668>
- Engelhard, G., Jr., & Wang, J. (2024). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9781003458746>

- Forgasz, H., Hall, J., & Robinson, T. (2024). Evaluating pre-service teachers' statistical literacy capabilities. *Mathematics Education Research Journal*, 36(1), 231–258. <https://doi.org/10.1007/s13394-022-00438-6>
- Friedrich, A., Schreiter, S., Vogel, M., Becker-Genschow, S., Brünken, R., Kuhn, J., Lehmann, J., & Malone, S. (2024). What shapes statistical and data literacy research in K-12 STEM education? A systematic review of metrics and instructional strategies. *International Journal of STEM Education*, 11(1), 58. <https://doi.org/10.1186/s40594-024-00517-z>
- Geiger, V., & Schmid, M. (2024). A critical turn in numeracy education and practice. *Frontiers in Education*, 9, 1363566. <https://doi.org/10.3389/feduc.2024.1363566>
- Hudson, N. W., Anusic, I., Lucas, R. E., & Donnellan, M. B. (2020). Comparing the reliability and validity of global self-report measures of subjective well-being with experiential day reconstruction measures. *Assessment*, 27(1), 102–116. <https://doi.org/10.1177/1073191117744660>
- Koga, S. (2022). Characteristics of statistical literacy skills from the perspective of critical thinking. *Teaching Statistics*, 44(2), 59–67. <https://doi.org/10.1111/test.12302>
- Li, J., Pan, D., Zhao, Y., & Long, T. (2026). From video to analytics: A design-based study on fostering critical collective reflection for pre-service science teachers. *Journal of Science Education and Technology*. <https://doi.org/10.1007/s10956-026-10356-y>
- Liu, Y., Wang, J., Zhang, Z., Wang, J., Luo, T., Lin, S., Li, J., & Xu, S. (2024). Development and validation of an instrument for measuring civic scientific literacy. *Disciplinary and Interdisciplinary Science Education Research*, 6(1), 6. <https://doi.org/10.1186/s43031-023-00092-3>
- Nima, A. A., Cloninger, K. M., Persson, B. N., Sikström, S., & Garcia, D. (2020). Validation of subjective well-being measures using item response theory. *Frontiers in Psychology*, 10, 3036. <https://doi.org/10.3389/fpsyg.2019.03036>
- Ozturk, A. B., & Baki, A. (2024). Models of conceptualizing and measuring statistical knowledge for teaching: A critical review. *Pedagogical Research*, 9(1), em0178. <https://doi.org/10.29333/pr/14031>
- Pretorius, T. B. (2023). The Perceived Stress Scale is essentially unidimensional: Complementary evidence from ancillary bifactor indices and Mokken analysis. *Acta Psychologica*, 241, 104058. <https://doi.org/10.1016/j.actpsy.2023.104058>
- Ribeiro Santiago, P. H., Nielsen, T., Smithers, L. G., Roberts, R., & Jamieson, L. (2020). Measuring stress in Australia: Validation of the Perceived Stress Scale (PSS-14) in a national sample. *Health and Quality of Life Outcomes*, 18(1), 100. <https://doi.org/10.1186/s12955-020-01343-x>
- Rizbudiani, A. D., Jaedun, A., Rahim, A., & Nurrahman, A. (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables (SLETV). *Al-Jabar: Jurnal Pendidikan Matematika*, 12(2), 399–412. <https://doi.org/10.24042/ajpm.v12i2.9939>
- Schreiter, S., Friedrich, A., Fuhr, H., Malone, S., Brünken, R., Kuhn, J., & Vogel, M. (2024). Teaching for statistical and data literacy in K-12 STEM education: A systematic review on teacher variables, teacher education, and impacts on classroom practice. *ZDM – Mathematics Education*, 56(1), 31–45. <https://doi.org/10.1007/s11858-023-01531-1>
- Schüller, K. (2022). Data and AI literacy for everyone. *Statistical Journal of the IAOS*, 38(2), 477–490. <https://doi.org/10.3233/SJI-220941>
- Shahat, M. A., Boone, W. J., Al-Alawi, K. A., & Al-Balushi, S. M. (2024). Rasch analysis in developing a questionnaire to measure self-efficacy beliefs of Omani preservice science teachers for teaching through engineering design. *Humanities and Social Sciences Communications*, 11(1), 1528. <https://doi.org/10.1057/s41599-024-04087-x>
- Shaw, A., Kapnek, M., & Morelli, N. A. (2021). Measuring creative self-efficacy: An item response theory analysis of the Creative Self-Efficacy Scale. *Frontiers in Psychology*, 12, 678033. <https://doi.org/10.3389/fpsyg.2021.678033>
- Shimizu, I., & Ferreira, J. C. (2023). Losing your fear of using R for statistical analysis. *Jornal Brasileiro de Pneumologia*, 49(3), e20230212. <https://doi.org/10.36416/1806-3756/e20230212>
- Sinclair, S., Hack, T. F., MacInnis, C. C., Jaggi, P., Boss, H., McClement, S., Sinnarajah, A., & Thompson, G. (2021). Development and validation of a patient-reported measure of compassion in healthcare: The Sinclair Compassion Questionnaire (SCQ). *BMJ Open*, 11(6), e045988. <https://doi.org/10.1136/bmjopen-2020-045988>

- Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, 46(3), 604–617. <https://doi.org/10.1080/09638288.2023.2169772>
- Umbach, G. (2022). Statistical and data literacy in policy-making. *Statistical Journal of the IAOS*, 38(2), 445–452. <https://doi.org/10.3233/SJI-220962>
- Welter, V. D. E., Emmerichs-Knapp, L., & Krell, M. (2023). Are we on the way to successfully educating future citizens?—A spotlight on critical thinking skills and beliefs about the nature of science among pre-service biology teachers in Germany. *Behavioral Sciences*, 13(3), 279. <https://doi.org/10.3390/bs13030279>
- Wu, X., Xu, T., & Zhang, Y. (2023). Research on the data analysis knowledge assessment of pre-service teachers from China based on cognitive diagnostic assessment. *Current Psychology*, 42(6), 4885–4899. <https://doi.org/10.1007/s12144-021-01836-y>
- Zhunissova, S., Zhussupova, L., Abyzbekova, G., & Islambekuly, B. (2026). Learning to trust experimental data: A validation-based approach to experiment design in pre-service chemistry teacher education. *Chemistry Education Research and Practice*, 27(3), 1107–1122. <https://doi.org/10.1039/D6RP00078A>